algorithmica
re-engineering it talent

BIGDATA Processing with Hadoop (HIVE)

Problem 1: Game and Player Analytics

The domain that we will model is a Game Center system that stores information on games, players, and their participations. A domain description:

- A **Game** consists of an id, name, publisher name, release date, and rating.
- A **Player** has an id, firstName, lastName, birthDate, and gender
- A **PlayerGame** consists of playerId, gameId and score. For each game played, a player has an integer score.

You can find sample data files games.txt, players.txt and player_games.txt at algorithmica github repository under datasets branch. Here are CREATE TABLE definitions for the files. Note that we are using EXTERNAL tables.

```
CREATE EXTERNAL TABLE Games (gid INT, gname STRING, pubname STRING,
releaseDate STRING, rating DOUBLE)
ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t'
LOCATION '/user/cloudera/problem1/games';

CREATE EXTERNAL TABLE Players (pid INT, fname STRING, lname STRING,
bdate STRING, gender STRING)
ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t'
LOCATION '/user/cloudera/problem1/players';

CREATE EXTERNAL TABLE PlayerGames (pid INT, gid INT, score INT)
ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t'
LOCATION '/user/cloudera/problem1/playergames';
```

Write HIVE queries to perform the following tasks:

Task #1 List all Games

Write a HiveQL query that will list all the game records. There should be 100 records printed.

Task #2 Find a Game by its Id

Write a HiveQL query that given some game id (hard-coded constant) will return the game record if found.

BIGDATA Processing with Hadoop (HIVE)

Task #3 Players 18 and over

Write a HiveQL query that lists only the players that are 18 and over. The output should be sorted by age ascending. You can do this question without a UDF if you look at the date function support of Hive.

Task #4 Number of Players per Game

Write a HiveQL query that will calculate the number of players per game. The output does not have to be sorted.

Task #5 Given a Game Id - List the Top 10 Scores for the Game

Write a HiveQL query that will output the top 10 scores in descending order for a given game id.

Task #6 Players in Common

Write a HiveQL query that will output pairs of game ids and the number of players they have in common. For instance, if game X and game Y have 2,000 players in common (play both games), then output X, Y, 2000. The data does not have to be sorted.

Task #7 Count Top Players for Each Game

Write a HiveQL query that will list all games along with the count of the number of players in each game with a score over 98,000. If a game does not have a player with a score over 98,000, it should still appear in the output with a count of 0.

Task #8 List All Players with Certain Properties

Write a HiveQL query that will list all players that either have a score in some game over 90,000 or play a game published by 'Electronic Arts'.

Task #9 List Most Popular Games for Each Publisher by Gender

Write a HiveQL query that for each publisher will list two records. The first will be the publisher id, "female", and the maximum number of women that play one of its games. The second row will be the publisher id, "male", and the maximum number of men that play one of its games. Note that the games may be different.

Task #10 Show Game Breakdown by Gender

For each game, show the total number of players, the total number of female and male players, and the percentage of female and male players.

BIGDATA Processing with Hadoop (HIVE)

Problem 2: Bigram Analytics

Bigrams are simply sequences of two consecutive words. For example, the previous sentence contains the following bigrams: "Bigrams are", "are simply", "simply sequences", "sequence of", etc. Use the bible ascii file for test data(we used it in MapReduce Assignment).

a) Write a PIG script/MR program to compute all bigrams with frequencies of given text file.

b) Create a HIVE external table that refers to the output data generated from previous question.

c) Write HIVE queries to answer following questions:

i. How many unique bigrams are there?
ii. List the top ten most frequent bigrams and their counts.
iii. What is the cumulative frequency of the top ten bigrams?
iv. How many bigrams appear only once?

Problem 3: MovieLens Data Analytics

GroupLens Research has collected and made available rating data sets from the MovieLens web site (http://movielens.umn.edu). The input consists of a series of lines, each containing a userID movieID, rating and timestamp. The dataset can be found here:
http://files.grouplens.org/datasets/movielens/ml-100k/u.data

You are required to write efficient HIVE queries to find the following information:

a) The average ratings of each movie
b) k most popular movies of all time
c) k most popular movies for each year